H I L G A R D I A

A Journal of Agricultural Science Published by the California Agricultural Experiment Station

VOL. 16 FEBRUARY, 1944 No. 1

BIASES ENCOUNTERED IN LARGE-SCALE YIELD TESTS[1,2]

O. C. RIDDLE[3] and G. A. BAKER[4]

INTRODUCTION

CERTAIN DIFFICULTIES in interpreting the results of the ordinary analysis of variance when applied to yield tests of genetically similar wheat strains, derived through backcrossing, prompted a critical examination of the data. These data indicated a bias inherent in any large-scale experiment that imposes an inflexible design upon a soil whose fertility may fluctuate markedly within short distances. Because of this bias, the usual analysis of variance tends to overestimate significance grossly when the number of varieties is large and the productivity levels of the soil change rapidly and erratically (as at Davis). One may briefly describe the bias by saying that the inflexible design of the experiment tends to subtract too little from the naturally high-yielding plots and too much from the naturally low-yielding plots in attempting to correct for differences in soil productivity. This has a spreading effect on the part of the variation that is labeled "varietal differences" and thus causes a serious overestimation of significance.

In a conventional analysis of variance, the natural variation of an experiment is arbitrarily partitioned into categories according to a preconceived mathematical model. These categories are labeled "variation due to varieties," "variation due to soil productivity," and so on. If the experiment is in exact accord with the model, the labels are accurate. If the experiment is not as called for by the model, the labels are misleading: for instance, the category labeled "variation due to varieties" may contain some of the variation due to soil productivity.

A mathematical model may be pleasing and beautiful to its creator or users. If, then, nature does not conform to the model, workers having limited first-hand experience with the vagaries of biological material may even feel that nature has erred and should be corrected. Baten, Northam, and Yeager (*2*)[5],

[1] Received for publication May 14, 1943.

[2] Results of a coöperative study conducted by the United States Bureau of Plant Industry Division of Cereal Crops and Diseases, and the University of California Division of Agronomy and Division of Mathematics and Physics, Davis, California.

[3] Instructor in Agronomy and Junior Agronomist in the Experiment Station. Formerly Agent, United States Department of Agriculture.

[4] Assistant Professor of Mathematics and Assistant Statistician in the Experiment Station.

[5] Italic numbers in parentheses refer to "Literature Cited" at the end of this paper.

for example, in a recent issue of *Journal of the American Society of Agronomy,* throw out the observed yields on two strains of tomatoes because those yields are not in accord with their mathematical model. The observed yields are replaced by computed yields based on the other observed yields. There is another strong temptation: a model that has proved useful in a restricted realm may be unquestioningly extrapolated to new and unexplored situations. This has been done by workers in all branches of science. One main purpose of this paper is to point out the danger of such extrapolation.

To construct an adequate mathematical model of some portion of nature, one must have extensive and accurate data on all situations to which the model is to apply. Such data are not now available. This paper aims to stimulate the collection of data that will provide for an improved model for yield trials. It may help to show that such data are necessary.

GENETIC RELATIONSHIP OF MATERIAL

As pointed out by Suneson and his co-workers (*17*), 182 F_3 strains of the pedigree (Martin × White Federation[6]) × (Hope × White Federation[5]) and 157 F_3 strains of (Martin × Baart[7]) × (Hope × Baart[5]) were bulked to produce White Federation 38 and Baart 38, respectively. These two new wheat varieties have been shown (*17*) to be essentially like their prototypes except for the incorporated resistance to bunt, or stinking smut (*Tilletia tritici*), and to stem rust (*Puccinia graminis* var. *tritici*). From the strains mentioned above, selections were made at random for the yield tests reported in this paper.

An understanding of the genetic relationship of these strains is important in consideration of the results of this test and in the extension of the implications therefrom to other methods of testing similar or unrelated material.

As Briggs (*3, 4*) has pointed out, "the proportion of homozygous individuals in any backcross generation is the same as would result from an equal number of selfed generations." This proportion may be calculated from the equation

$$\text{per cent homozygosity} = \left(\frac{2^m - 1}{2^m}\right)^n \times 100, \qquad (1)$$

where m is the number of generations of backcrossing and n is the number of heterozygous factor pairs in the original cross.

As Jones (*8*, p. 331) explains, "the proportion of complete homozygotes to the different classes of heterozygotes in any generation" can be obtained by expanding the binomial

$$[1 + (2^m - 1)]^n, \qquad (2)$$

where m and n are as above.

If we assume that the wheat parents used to develop the strains under test differ by 21 factor pairs (that is, a 1-factor difference on each of their chromosome pairs) and that the genotypes have been randomly sampled in the course of backcrossing, we can approximate the degree of homozygosity of the wheat strains in question. On the basis of these two assumptions, we can show by means of equations 1 and 2 that in the F_3 of the sixth backcross, 82.3 per cent of the plants are homozygous for the recurrent parent genotype, and that an additional 13.5 per cent differ by only 1 factor.

Admittedly, the parents differ by more than 21 factor pairs; but calculations accounting for all factor differences and the effects of linkage are not

possible. Certain considerations suggest, however, that the estimates of homozygosity given above may approach the true situation. As the fraction $\frac{2^m - 1}{2^m}$ (from equation 1) approaches unity, n (the number of heterozygous factor pairs) will change that value but little. Through the mechanism of crossing over and random assortment of chromosomes, repeated backcrossing facilitates recovery of the complete chromosome complement of the recurrent parent except for a small segment from the nonrecurrent parent on which the gene being selected for is located. In developing these strains, rigid selection toward the recurrent parent phenotype was practiced in early backcross generations; this hastens return to the recurrent parental genotype and reduces the unfavorable effects of linkage.

Actually, under the conditions of these tests the strains were morphologically indistinguishable, except strain 1441, which averaged 2 inches taller than the others. No other differences in growth habit were observed at any stage, and there was no detectable difference in reaction to any disease among the strains, although the original Baart and White Federation grown with them were attacked by rust.

Although the strains are not genetically identical, the degree of similarity is clearly such as to require a critical test of significance to detect any possible differences in yielding ability.

METHODS AND DESIGN

Twenty-nine strains of Baart 38 and thirty-four strains of White Federation 38 were chosen at random for yield testing. These, together with the disease-susceptible prototypes, were set up in separate experiments in each of the years 1939 and 1940. The design employed for these tests was a modified Latin square suggested by "Student" (*16*) and by Snedecor (*13,* p. 38) and used extensively by Pope (*11*) and others (*15, 21*). There were five replications divided into five columns superimposed upon and situated at right angles to the replications. The strains were grown in single 16-foot rows. They were completely randomized except for the double restriction that each strain occurred once (and only once) in each replication and each column. The same randomization but totally different fields were used for the tests in the two years. The high and low extremes from the 1939 and 1940 tests of both the Baart 38 and the White Federation 38 strains, each with its susceptible prototype, were tested in separate 6×6 Latin squares in 1941 using single 16-foot-row plots.

The data were analyzed by use of the analysis of variance, testing for significance by F (*14,* p. 184). Least significant differences above and below the general mean were established by the use of t (*14,* p. 58) for the appropriate degrees of freedom. The values of $\frac{\text{range}}{S.E.}$ expected from random sampling in a normal homogeneous population of sample size N were obtained from Snedecor (*14,* p. 89).

Yates (*19*) has criticized the modified Latin square as being subject to a biased estimate of error. Repeated reference to this criticism in the literature (*5, 20*) condemns the design in favor of the more complex incomplete block

designs. The present writers do not defend the modified Latin square design nor advocate its use; but they feel that information can be gained from the data in these experiments in which the modified Latin square design was used.

There is evidence that the error estimates in these experiments are not biased in the sense of Yates's (*19*) criticism. The F values (table 4) due to strain differences are greater than required for the 1 per cent level of significance in all cases for the 1939 and 1940 modified Latin square tests. If these F values are large only because of a biased estimate of error, then the estimates of error variance must be considered *too small in all cases.* We may compare certain variances as a test of Yates's bias. The degrees of freedom for the Baart 38 tests, for instance, may be set out from what Fisher (*7*) calls a topographical standpoint as:

Between plots	replications	4
	columns	4
	replications × columns	16
Within plots		125
Total		149

If the interaction variance of replications × columns is the same as the within-plot variance, then no Yates's bias can exist. A comparison of these variances for the 1939 and 1940 experiments in table 1 shows no significant nor consistent differences when tested by F and therefore no indication of Yates's bias.

TABLE 1

COMPARISON OF VARIANCES INDICATING NO YATES'S BIAS IN THE DATA

Experiment and year	Interaction variance	Within-plot variance	F	F, 5 per cent
Baart 38—1939	5,361	3,861	1.39	1.72
Baart 38—1940	5,638	5,905	1.05	2.04
White Federation 38—1939	3,096	3,166	1.02	2.04
White Federation 38—1940	7,479	5,404	1.38	1.71

Yates's bias, if it exists, seems to be sometimes in one direction and sometimes in the other, hence resembles the biases considered by Welch (*18*). Such a bias can be expected to balance out in the long run; one could allow for it by slightly changing the probability levels at which significance is accepted or rejected.

EXPERIMENTAL RESULTS

Yield data for the Baart 38 and White Federation 38 component strains tested in 1939, 1940, and 1941 are given in tables 2 and 3, respectively. The several strains are listed in descending order of average yields for 1939 and 1940 combined. Row numbers per replication are listed for each strain in recording the 1939 and 1940 results, to facilitate additional treatment of the data if desired. Table 4 summarizes the analysis of variance for each experiment, and gives pertinent statistical constants.

Indicated Significances.—In all 1939 and 1940 tests, F values due to differences in mean yields of strains exceed those required at the 1 per cent level of significance (see table 4). Standard errors were calculated; and minimum

TABLE 2

Yields and Row Numbers for 1939–1940 Modified Latin Square Yield Trials, and Mean Yields for 1941 Latin Square Test of Baart and of Randomly Selected Component Strains of Baart 38

Strain no.	Yield in grams per 16-foot row and row number per replication: 1939–1940															Strain mean 1939	Strain mean 1940	1939–1940 combined mean	Strain mean 1941
	Replication I			Replication II			Replication III			Replication IV			Replication V						
	Row no.	1939 yield	1940 yield	Row no.	1939 yield	1940 yield	Row no.	1939 yield	1940 yield	Row no.	1939 yield	1940 yield	Row no.	1939 yield	1940 yield				
5037	54	575	633	60	540	477	68	790	782	39	445	628	49	560	630	582*	630.0*	606.0*	401.0
3061	44	500	606	45	440	590	52	655	782	68	450	480	62	660	786	541	648.8*	594.9*	419.5
2725	42	450	604	63	585	523	54	740	755	59	390	543	50	525	612	538	607.4	572.7	.....
4673	61	540	522	41	410	485	45	525	675	65	530	537	54	750	738	551	591.4	571.2	.....
6061	57	565	594	53	615	515	66	745	550	50	345	535	43	505	696	555	578.0	566.5	.....
4869	53	620	557	68	630	572	57	645	595	43	340	537	45	590	538	565*	559.8	562.4	.....
4630	59	560	522	64	530	302	44	525	723	53	570	572	46	550	712	547	566.2	556.6	.....
4621	51	475	525	50	480	590	63	700	557	60	420	503	39	625	622	540	559.4	549.7	.....
3541	46	395	600	55	585	539	62	665	583	64	450	502	42	540	570	527	558.8	542.9	.....
3257	68	570	587	49	480	457	55	700	590	44	340	588	59	585	505	535	545.4	540.2	.....
3177	45	440	579	42	480	532	65	655	513	61	470	586	55	570	563	523	554.6	538.8	.....
6045	56	505	504	66	690	378	50	545	554	41	390	548	58	660	612	558	519.2	538.6	.....
4641	60	545	521	54	655	480	40	560	628	45	375	498	63	635	476	554	520.6	537.3	.....
5129	62	450	660	51	410	488	47	540	645	40	375	614	65	500	674	455†	616.2*	535.6	.....
5217	55	520	646	47	450	548	41	530	570	62	530	520	68	595	585	525	573.8	531.4	.....
2685	40	495	630	57	570	392	64	575	535	55	475	580	47	485	567	520	540.8	530.4	.....
5493	64	420	575	56	550	491	59	670	497	49	410	430	41	490	725	508	543.6	525.8	.....
3593	47	365	465	61	540	434	42	405	550	54	545	586	66	600	716	491	550.2	520.6	.....
2645	39	420	548	52	530	385	67	685	465	48	375	603	60	510	660	504	532.2	518.1	.....
3509	66	485	565	43	390	574	51	475	583	46	330	610	61	625	526	461	571.6	516.3	.....
4761	67	570	553	46	365	593	61	670	450	42	335	554	52	515	549	491	539.8	515.4	.....
3681	48	405	508	65	510	500	39	565	452	58	435	592	51	495	687	482	547.8	514.9	.....
4589	58	565	494	39	515	493	48	435	657	67	475	540	56	520	435	502	523.8	512.9	.....
4025	50	410	568	40	445	508	58	585	579	51	330	545	67	620	482	478	536.4	507.2	.....
2765	43	390	473	67	530	528	49	360	557	52	415	564	57	605	600	460†	544.4	502.2	.....
Baart	65	480	555	62	545	372	43	475	538	56	430	545	48	535	533	493	508.6	500.8	365.7
2697	41	390	526	48	390	555	60	545	560	66	450	478	53	550	532	465	530.2	497.6	.....
4817	52	445	570	44	380	580	46	525	527	57	415	425	64	415	587	436†	537.8	486.9†	.....
5317	63	455	438	58	610	422	56	565	430	47	315	437	40	555	497	500	444.8†	472.4†	431.3
3769	49	435	420	59	550	430	53	530	504	63	445	357	44	515	485	495	439.2†	467.1†	390.5
Baart 38‡	..	...	...	..	...	...	..	...	...	..	...	...	..	...	...	...	.....	.....	377.2‡

* Significantly higher than the general mean. The general mean was 512.7 in 1939, 550.7 in 1940, and 531.7 for the combined yields.
† Significantly lower than the general mean.
‡ The mixture, Baart 38, was not grown in 1939 and 1940 since the object was to determine if Baart 38 could be separated into parts that yield differently.

TABLE 3

Yields and Row Numbers for 1939–1940 Modified Latin Square Yield Trials, and Mean Yields for 1941 Latin Square Test of White Federation and of Randomly Selected Component Strains of White Federation 38

Strain no.	Yield in grams per 16-foot row and row number per replication: 1939–1940															Strain mean 1939	Strain mean 1940	1939–1940 combined mean	Strain mean 1941
	Replication I			Replication II			Replication III			Replication IV			Replication V						
	Row no.	1939 yield	1940 yield	Row no.	1939 yield	1940 yield	Row no.	1939 yield	1940 yield	Row no.	1939 yield	1940 yield	Row no.	1939 yield	1940 yield				
2029	26	485	610	17	665	748	30	650	710	15	560	629	3	640	665	600	672.4*	636.2*	440.0
153	4	580	705	33	640	598	24	650	623	12	670	528	17	560	630	620	616.8	618.4*	397.7
1341	16	445	700	6	705	573	32	710	721	25	540	492	9	655	625	611	622.2	616.6*	.....
2453	33	500	495	27	610	654	2	635	598	13	730	640	18	595	692	614	615.8	614.9	.....
1949	25	515	600	7	695	575	15	690	597	16	630	525	36	565	725	619	604.4	611.7	.....
2249	30	535	505	20	680	524	6	760	672	11	665	552	29	605	600	649*	570.6	609.8	.....
873	12	635	585	32	640	552	22	700	621	6	625	576	26	495	649	619	596.6	607.8	.....
2185	28	410	630	34	605	611	4	655	675	18	515	665	12	650	652	567	646.6*	606.8	.....
1809	23	415	603	31	595	712	20	660	583	10	540	625	7	625	678	567	640.2*	603.6	.....
1673	19	460	622	30	625	498	8	705	650	9	630	633	27	515	647	587	610.0	598.5	.....
229	6	600	750	23	575	639	31	670	517	14	575	617	20	440	594	572	623.4	597.7	.....
1825	24	425	525	22	670	602	35	610	700	4	550	616	13	660	594	583	607.4	595.2	.....
2357	31	575	445	24	610	588	13	850	539	20	530	578	6	715	508	656*	531.4	593.7	.....
1785	22	470	585	13	805	456	28	665	617	33	480	652	5	560	630	596	588.0	592.0	.....
257	7	550	517	9	630	550	27	590	667	17	505	670	35	610	620	577	604.8	590.9	.....
1441	17	485	590	11	720	581	5	710	614	31	440	612	23	530	580	577	595.4	586.2	.....
313	8	545	585	18	625	622	23	805	505	32	460	623	14	550	533	597	573.6	585.3	.....
2193	29	555	515	14	690	600	21	770	532	35	515	514	8	550	601	616	552.4	584.2	.....
1617	18	445	609	25	610	603	10	700	629	2	405	635	30	505	698	533†	634.8*	583.9	.....
2597	35	500	590	10	630	525	25	760	552	7	600	547	16	590	526	616	548.0	582.0	.....
2061	27	450	555	36	550	579	16	670	583	8	630	542	10	530	729	566	597.6	581.8	.....
33	3	490	490	12	750	436	34	720	674	19	460	614	28	475	672	579	577.2	578.1	.....
1185	14	515	600	21	565	607	33	605	522	23	505	527	4	630	622	564	575.6	569.8	.....
357	9	520	498	4	655	513	17	630	560	34	525	588	24	545	642	575	560.2	567.6	.....
1741	21	465	587	5	645	493	11	760	575	36	450	618	25	525	538	569	562.2	565.6	.....
957	13	560	475	3	610	494	29	565	563	22	530	585	34	570	680	567	559.4	563.2	.....
437	10	560	505	35	640	567	7	710	527	28	435	518	22	535	605	576	544.4	560.2	.....
White Federation	36	410	420	29	595	617	9	785	530	5	575	520	21	565	573	586	532.0	559.0	394.8
2417	32	490	427	8	675	493	19	670	609	29	340	661	15	570	640	549	566.0	557.5	378.2
2457	34	530	455	15	650	597	26	695	603	3	550	379	19	580	521	601	511.0†	556.0	.....
569	11	500	488	16	620	554	3	730	538	27	405	525	31	595	570	570	535.0	552.5	.....
1717	20	410	520	2	570	564	14	760	564	24	400	545	32	535	654	535†	569.4	552.2	.....
5	2	515	555	19	600	589	12	760	557	26	485	529	33	485	427	569	531.4	550.2	.....
1309	15	435	580	28	485	456	18	615	586	30	400	491	2	465	565	480†	535.6	507.8†	379.2
169	5	500	385	26	630	337	36	640	420	21	515	355	11	635	429	584	385.2†	484.6†	.....
White Federation 38‡	..	...	...	..	...	...	..	...	...	..	...	...	..	...	...	...	.....	.....	405.7‡

* Significantly higher than the general mean. The general mean was 584.2 in 1939, 577.1 in 1940, and 580.6 for the combined yields.
† Significantly lower than the general mean.
‡ The mixture, White Federation 38, was not grown in 1939 and 1940 since the object was to determine if White Federation 38 could be separated into parts that yield differently.

TABLE 4

SUMMARY OF VARIANCE ANALYSIS AND STATISTICS OF BAART 38 AND WHITE FEDERATION 38 STRAINS TESTED FOR YIELD, 1939–1941

Experiment and source of variation	Degrees of freedom	Sum of squares	Mean square	*F* value		
				Actual	5 Per cent	1 Per cent
a. Baart 38 strains—1939:						
Between means of replicates....	4	524,648	131,162			
Between means of columns.....	4	367,251	91,813			
Between means of strains.......	29	200,139	6,901	2.10	1.57	1.89
Error..........................	112	368,291	3,288			
Total..........................	149	1,460,329				
b. Baart 38 strains—1940:						
Between means of replicates....	4	202,727	50,682			
Between means of columns.....	4	40,766	10,192			
Between means of strains.......	29	283,937	9,791	2.02	1.57	1.89
Error..........................	112	544,176	4,859			
Total..........................	149	1,071,606				
c. Baart 38 strains—1941:						
Between means of replicates...	5	25,432	5,086			
Between means of columns.....	5	16,783	3,357			
Between means of strains.......	5	18,700	3,740	0.71	2.71	4.10
Error..........................	20	105,850	5,292			
Total..........................	35	166,765				
d. White Federation 38 strains—1939:						
Between means of replicates....	4	894,616	223,654			
Between means of columns.....	4	246.805	61,701			
Between means of strains.......	34	186,975	5,499	2.15	1.55	1.85
Error..........................	132	337,509	2,557			
Total..........................	174	1,665,905				
e. White Federation 38 strains—1940:						
Between means of replicates....	4	75,467	18,867			
Between means of columns.....	4	9,565	2,391			
Between means of strains.......	34	438,644	12,901	3.46	1.55	1.85
Error..........................	132	491,588	3,724			
Total..........................	174	1,015,264				
f. White Federation 38 strains—1941:						
Between means of replicates....	5	80,757	16,151			
Between means of columns.....	5	4,422	884			
Between means of strains.......	5	15,430	3,086	1.02	2.71	4.10
Error..........................	20	60,613	3,031			
Total..........................	35	161,222				

Statistics	Experiment					
	a	b	c	d	e	f
1. General mean (gms. per 16 ft. row)........................	512.7	550.7	397.5	584.2	577.1	399.2
2. *S.E.* of a single plot (gms.)................................	57.3	69.7		50.6	61.0	
3. Least significant difference at 5 per cent level between any strain mean and general mean (gms.)............	51.4	62.6		45.2	54.6	
4. Range in strain means/*S.E.* of a strain mean:						
Observed..	5.7	6.7		7.8	10.5	
Expected (approximate)...............................	4.1	4.1		4.1	4.1	
5. Coefficient of variability (per cent).......................	5.0	5.7		3.9	4.7	

significant differences, at the 5 per cent level, above and below the general mean were established. The general mean yield of all strains was used as the reference point for judging significance because, if significant differences did exist, the strains higher in yield than the general mean of all strains would be of greatest agronomic value.

As indicated in tables 2 and 3, certain strains were found to differ significantly from the general mean in both 1939 and 1940 and in the combined 1939 and 1940 results. In addition, two criteria suggest that the observed mean differences are not of the order expected from random sampling in a homogeneous population: first, the high values of F; second, the high values of range in mean yields divided by standard error of the mean (table 4).

Difficulties of Interpretation.—Considering the genetic relationship of the strains tested, we might expect that their mean yields would not be significantly different. Yet, as mentioned above, significant differences above and below the general mean are indicated, and they are such as to deserve consideration agronomically. Furthermore, if differences do exist, the strains might be expected to maintain somewhat the same relative positions in repeated tests. Certainly such characters as growth habit at any stage, plant height, or date of maturity showed no observable difference that would suggest a differential response to environment. Moreover, of the strains indicated as significantly different from the general mean in tables 2 and 3, *only one* strain, 5037 (a Baart 38 strain), holds the same position for the two years, 1939 and 1940, in having a yield significantly greater than the general mean. Two strains, namely 5129 (a Baart 38 strain) and 1617 (a White Federation 38 strain), reversed their position from *significantly lower* in 1939 to *significantly higher* than the general mean in 1940. All other strains that were significantly different in one or the other of the two years were not significantly different in the alternate year. As will be observed in tables 2 and 3, the Baart and White Federation prototypes did not differ significantly from the general mean in 1940 even though *they were attacked by stem rust.*

So far in the analysis, two reactions have been indicated that are not in accord with expectation if the genetic relationship of the strains agrees with that expressed under the section on "Genetic Relationship of Material." These reactions are, namely, significant differences in yield, and failure to exhibit comparable relative yield responses in two different years.

The correlation of one year's yields with another year's depends upon the variance of the true yields and upon the experimental-error variance. If for the two years the variances of the true yields are the same, if the strains maintain their relative positions, and if the error variance is the same for every plot, then the expected correlation between the yields for the two years is a value equal to the variance of the true yields divided by the quantity true-yield variance plus strain-mean variance; hence it is less than 1. Conceivably, the spread among the true yields might be large enough to insure detection four times out of four, and yet small enough relative to the error of the experiment so that the correlation between yields for two years might be small. Neyman's tables[6] for the probability of failing to detect differences between

[6] Supplied in manuscript form by Professor J. Neyman of the Statistical Laboratory, University of California, Berkeley.

yields when they actually exist show that this situation is not possible for these data.

The 1941 6×6 Latin square yield test, sampling the 1939 and 1940 high-low extremes, constituted a more critical attempt to determine whether the strains actually differ in yielding ability. If true yield differences exist, they should certainly be represented in those extremes indicated by previous tests to be significantly different. As shown in the analysis of variance applied to the 1941 data (table 4), the *F values due to differences in mean yields of previously indicated high-low strains are not significant.* The results of this test do not indicate differences in yield of the strains tested. Neyman's tables indicate that this test was sufficient to detect, with a probability greater than 0.8, differences of the order necessary to account for the observed F values if the usual mathematical model is assumed.

Considering these difficulties, we may well examine critically the statistical methods used and the assumptions on which those statistics are based.

BASIC ASSUMPTIONS IN ANALYSIS OF VARIANCE

The derivation of the analysis-of-variance technique is based on four assumptions: (1) that the productivity levels of the plots assigned to any variety are independent of those assigned to any other; (2) that the estimates of individual plot yields are normally distributed about the "true" plot yield; (3) that the distribution of yield estimates for every plot has the same variance; (4) that in yield trials the productivity levels follow some prescribed law.

The work of Neyman (*10*) and McCarthy (*9*) relates to the first assumption. According to Neyman, the Latin square design may often indicate significance between hypothetical "varieties" in uniformity trials, partly because of unequal correlations between fertility levels of the plots assigned to the different "varieties." McCarthy shows further that these unequal correlations, when the varieties are tested in randomized blocks, may cause a serious overestimation of significance—that is, may indicate significance where none exists.

According to the second and third assumptions, the estimates of individual plot yields are normally distributed about the "true" plot yield with equal variances. In this connection the work of Baker (*1*) and Salmon (*12*) may be cited. Baker shows that the distribution of the estimates of a "true" plot yield is usually skewed one way or the other and that the variance of the distribution of estimates depends on the variations of the fertility within the plot. Sometimes the nonnormality of the distribution may be such as to cause a serious overestimation of significance. Baker shows further that adjacent plots of 15 square feet cannot be assumed to have the same fertility levels. Salmon and many others have discussed unequal variance as affecting the results of the analysis of variance.

The importance of assumption 4 is well recognized by some authorities (see Neyman [*10*]), but has been generally overlooked or deëmphasized by many workers concerned with yield trials. The present data show clearly the importance of the failure of conventional designs to prescribe a sufficiently flexible law for fertility levels.

Failure of the data to comply with any one of the assumptions on which the statistic is based may result in misleading or invalid conclusions.

CRITICAL EXAMINATION OF DATA

The nature and extent of the material tested and the relative simplicity of the design employed in these tests facilitate a critical examination of the data from the standpoint of the validity of the four assumptions mentioned above.

Residuals have been calculated and used in testing these data for the validity of the fundamental assumptions. Residuals are defined as plot yield plus twice the general mean minus the column mean minus the replication mean minus the variety mean.

We can roughly state the basic assumptions of the analysis-of-variance test in terms of residuals by saying that the residuals are normally distributed and that there is no pattern or system in the way in which they occur.

If we test by chi-square the hypothesis that the combined residuals of the 1939 and 1940 tests are normally distributed, then the resultant $P = 0.06$. Such a value of P indicates only a slight departure from normality. The distribution appears to be slightly peaked and positively skewed.

We should now examine the possibility that the residuals occur according to some plan or pattern.

If we consider that the residuals come at *random* from a normal population with a fixed standard deviation, then the six or seven residuals[7] occurring within a block (a block being the six or seven plots common to a given column and a given replication where the two cross at right angles) should be independent of the block-mean yield. These values *are not independent* for the 1939 and 1940 data. Thus when block-mean yields are correlated with block-mean residuals, the values of r for the Baart 38 strains in 1939 and 1940 are 0.23 and 0.24 respectively, and for the White Federation 38 strains for the same two years 0.23 and 0.54. Using the tables of David (*6*), one can calculate the probabilities of getting from a normal population for which the correlation is zero, correlation coefficients as high as the observed ones or higher. In samples of 25 these probabilities are 0.14, 0.13, 0.14, and 0.003 respectively. Let us compute, by the chi-square method of combining independent probabilities (*6*), the probability of the set of four observed values under the set of alternatives that the correlation coefficients of the sampled populations are unequal but all greater than zero. We find $P = 0.0006$. It is striking that the r values are the same for the three similar F values (see table 4), and that the much larger value of r occurs for the experiment with the exceedingly large F value.

Judging from the correlation between block-mean yields and block-mean residuals, too much has been subtracted from the poor plots and too little from the good plots in calculating the residuals. Hence, part of the variation in soil fertility has been assigned to the variation between varieties. That is, the design is too inflexible to take care of soil variation adequately from *one set of six or seven contiguous plots to another*. The result is a spreading effect on the strain means, and a tendency to indicate significance where none exists.

This correlation, furthermore, implies a parabolic relation between yield and the sum of squares of residuals.

Row totals (that is, a summation of yields of the plots occupying comparable

[7] Six in the experiments with Baart 38 strains and seven for White Federation 38 strains.

positions across all replications) show pronounced coincident peaks of fertility culminating at the sixteenth row for both years for the Baart 38 strains, though the experiments occupied different areas in the two years. These coincident productivity peaks explain why, in the analysis of variance, the same Baart 38 strain appeared significantly higher in yield in both years. The row totals of White Federation 38 strains show a similar peak in 1939, but no very definite peak in 1940.

We may briefly summarize the evidence from these studies and from the previously mentioned work of Neyman, McCarthy, Salmon, and Baker relating directly to the assumptions on which the analysis of variance is based. According to both Neyman (*10*) and McCarthy (*9*) there are some cases of unequally correlated levels of fertility of plots assigned to different varieties, and such correlation causes overestimation of significance when the analysis-of-variance technique is used. According to Baker (*1*), serious overestimation of significance may result from nonnormal distributions of yield estimates. According to Salmon (*12*) and others, the analysis of variance is invalid when the error variance differs from one part of the experiment to another. In the present work, correlation between fertility levels and residuals has been established. That is, the design has not prescribed a sufficiently flexible law of soil-productivity levels. This correlation means, not that residuals measured from their mean are more variable in one part of the experiment than in another, but that a bias in the residuals exists because soil productivity has been partially incorporated into strain differences. The bias to which we call attention is not attenuated as a cause of overestimation by the near identity of the strains tested. Certain conditions on soil-productivity levels, size of plot, and number of strains tested will lessen or make negligible the overestimation due to correlation between soil productivity and residuals.

APPLICATION TO OTHER YIELD-TESTING DESIGNS

So far as the authors are aware, no design now in use for testing large numbers of varieties is free from the danger of seriously overestimating significance when conditions are such that (1) fertility levels of plots assigned to different varieties are unequally correlated, (2) distribution of the estimates of a "true" plot yield is not normal, (3) the variances for different parts of the experiment are significantly different, and (4) an insufficiently flexible law is imposed on the productivity levels by an inadequate design. We believe, furthermore, that none of the conventional designs can adequately eliminate all these possible invalidating conditions under all conditions of testing. The lattice designs, now coming into vogue as the most efficient design for testing large numbers of varieties, have one admitted limitation : since varietal means are partially confounded with block effects, the use of these incomplete block designs may be cautioned against where large varietal differences are expected. In addition, the lattice designs seem to involve exactly the same difficulty experienced in these tests—namely, the danger of not subtracting the right amount from each plot yield or of not making the right "correction." They impose on the experiment a fixed formal framework, which may not be flexible enough to take care of spotted or abrupt changes in fertility levels.

Uniformity experiments are frequently recommended as preliminary steps

in determining the specific design best suited to test specific material in a given locality. Such tests, usually in operation for one year, or a very few years, cannot take into account the fluctuation of productivity levels from year to year as they are affected by changed environments, varying biological factors, tillage, and the like. They also provide no means of measuring the possible effect of complicating interactions when dissimilar rather than uniform material is under test.

We should not overlook the possibility that significance may be dangerously overestimated in any yield test of large numbers of varieties.

SUMMARY

An extensively used design, the modified Latin square, was used to test the comparative yielding ability of random selections from genetically similar component strains of Baart 38 and White Federation 38. Significant differences between strains were indicated in 1939. When the experiments were repeated in 1940, significant differences were again indicated, but with reversals from the previous year. When the results of the two years were combined, and the strains significantly greater and less than the general mean were again tested, these strains were found to be not significantly different. This last experiment was a small-scale Latin square experiment covering only a few strains. These facts prompted a critical examination of the data from the standpoint of the validity of the assumptions underlying the statistics used.

If the residuals in these experiments are examined, the block mean residual proves to be significantly and positively correlated with the block mean yield. Evidently, therefore, some of the difference in fertility levels of the plots has been assigned to strain differences. The result is a spreading effect on strain means, and a tendency to indicate significant differences in this large-scale experiment when, in fact, none exists.

Admittedly, small differences in yielding ability may actually exist between the strains tested in these experiments. Any such differences are masked, however, by the spotted variation in soil-fertility levels. Certainly the differences are not of the order of magnitude nor of the degree of significance indicated by the ordinary analysis of variance.

The demonstrated causes of overestimating significance in the analysis of variance are frequently assumed to be nonexistent. Indicated significant differences between "varieties" tested under conditions where any invalidating factors are operating should be viewed with skepticism.

LITERATURE CITED

1. BAKER, G. A.
1941. Fundamental distribution of errors for agricultural field trials. Natl. Math. Mag. **16**:7–19.

2. BATEN, W. D., J. I. NORTHAM, and A. F. YEAGER.
1941. Grouping of strains or varieties by use of a Latin square. Amer. Soc. Agron. Jour. **33**:616–22.

3. BRIGGS, FRED N.
1935. The backcross method in plant breeding. Amer. Soc. Agron. Jour. **27**:971–73.

4. BRIGGS, FRED N.
1938. The use of the backcross in crop improvement. Amer. Nat. **72**:285–92.

5. COX, G. M., R. C. ECKHARDT, and W. G. COCHRAN.
1939. The analysis of lattice and triple lattice experiments in corn varietal tests. Iowa Agr. Exp. Sta. Res. Bul. **281**:1–66.

6. DAVID, F. N.
1938. Tables of the correlation coefficient. Text 38 p. Tables 55 p. Biometrika Office, University College, London, Eng.

7. FISHER, R. A.
1935. Discussion *of*: Yates, F. Complex experiments. Roy. Statis. Soc. Jour. Sup. **2**:229–31.

8. JONES, DONALD FORSA.
1925. Genetics in plant and animal improvement. 568 p. John Wiley and Sons, Inc., New York, N. Y.

9. MCCARTHY, M. D.
1939. On the application of the Z-test to randomized blocks. Ann. Math. Statis. **10**:337–59.

10. NEYMAN, J.
1935. Statistical problems in agricultural experimentation. Roy. Statis. Soc. Jour. Sup. **2**:107–80.

11. POPE, O. A.
1936. Efficiency of single and double restrictions in randomized field trials with cotton when treated by the analysis of variance. Arkansas Agr. Exp. Sta. Bul. **326**:1–28.

12. SALMON, S. C.
1938. Generalized standard errors for evaluating bunt experiments with wheat. Amer. Soc. Agron. Jour. **30**:647–63.

13. SNEDECOR, GEORGE W.
1934. Calculation and interpretation of variance and covariance. 96 p. Collegiate Press, Inc., Ames, Iowa.

14. SNEDECOR, GEORGE W.
1938. Statistical methods. rev. ed. 388 p. Collegiate Press, Inc., Ames, Iowa.

15. STRINGFIELD, G. H., R. D. LEWIS, and H. L. PFAFF.
1941. The 1940 Ohio coöperative corn performance tests. Ohio Agr. Exp. Sta. Spec. Cir. **61**:1–30.

16. "STUDENT."
1931. Yield trials. *In:* Hunter, H. Baillière's Encyclopedia of Scientific Agriculture. Vol. 2, p. 1342–61. Baillière, Tindall, and Cox, London, Eng.

17. SUNESON, C. A., O. C. RIDDLE, and F. N. BRIGGS.
1941. Yields of varieties of wheat derived by backcrossing. Amer. Soc. Agron. Jour. **33**:835–40.

18. WELCH, B. L.
1937. On the Z–test in randomized blocks and Latin squares. Biometrika **29**:21–52.

19. YATES, F.
1935. Complex experiments. Roy. Statis. Soc. Jour. Sup. **2**:181–223, 243–47.

20. Zuber, M. S.
1942. Relative efficiency of incomplete block designs using corn uniformity trial data. Amer. Soc. Agron. Jour. **34**:30–47.

21. Zuber, M. S., and J. L. Robinson.
1941. The 1940 Iowa corn yield test. Iowa Agr. Exp. Sta. Bul. n.s. P19:519–93.

4½m-3,'44 (8828)